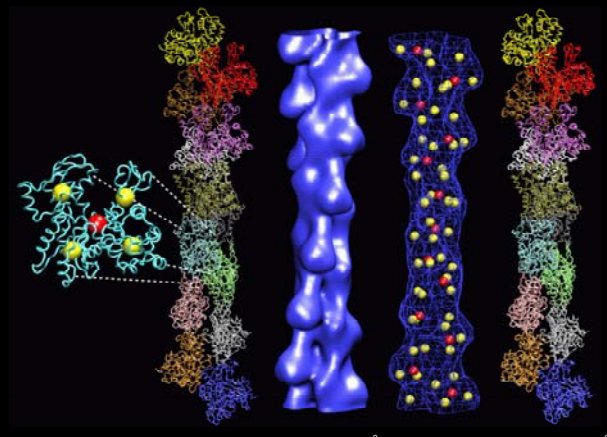
Point Cloud Registration with Anchor Point Matching

Willy Wriggers, Stefan Birmanns
biomachina.org

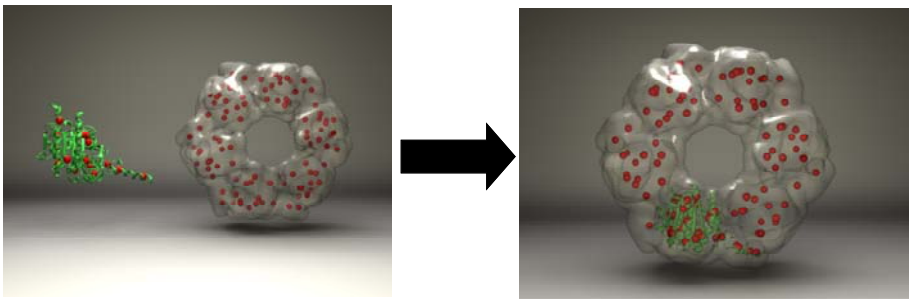
1998: "Simulated Markers"



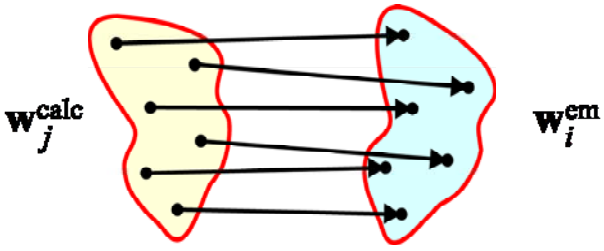
Actin filament: Reconstruction from EM data at 20Å resolution rmsd: 1.1Å

Willy Wriggers, Ronald A. Milligan, Klaus Schulten, and J. Andrew McCammon:
Self-Organizing Neural Networks Bridge the Biomolecular Resolution Gap.
J. Mol. Biol., 284:1247, 1998

1999-2007: Fast "Point Cloud" Fitting



Coarse-Grained Representations of Biomolecular Structure



Feature points (fiducials, landmarks), reduce complexity of search space

Useful for:

- Rigid-body fitting
- Flexible fitting
- Interactive fitting / force feedback
- Building of deformable models

Vector Quantization

Lloyd (1957) } Digital Signal Processing,
Linde, Buzo, & Gray (1980) } Speech and Image Compression.
Martinetz & Schulten (1993) Topology-Representing Network.

Encode data (in $\Re^{d=3}$) using a finite set $\{w_j\}$ (j=1,…,k) of *codebook vectors*.

Delaunay triangulation divides $\Re^3$ into k *Voronoi polyhedra* ("receptive fields"):

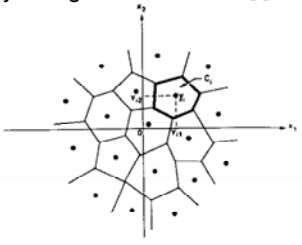
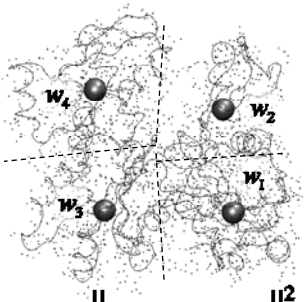


Fig. 3. Partitioning of two-dimensional space ($N = 2$) into $L = 18$ cells. All input vectors in cell C_i will be quantized as the code vector y_i. The shapes of the various cells can be very different.



Minimize encoding distortion error: $E = \sum_{i\,(\text{atoms, voxels})} \left\| v_i - w_{j(i)} \right\|^2 m_i$

Convergence and Variability

Q: How do we know that we have found the global minimum of E?

A: We don't (in general).

But we can compute the statistical variability of the $\{w_j\}$ by repeating the calculation with different seeds for random number generator.

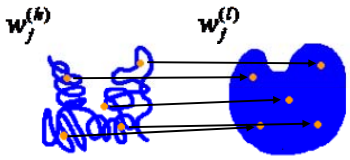
Codebook vector variability arises due to:
- statistical uncertainty,
- spread of local minima.

A small variability indicates good convergence behavior.
Optimum choice of # of vectors k: variability is minimal ("quality" of coarse-grained representation).

Single-Molecule Rigid-Body Docking

OLD 1.x style

Xtal structure



EM low res. data

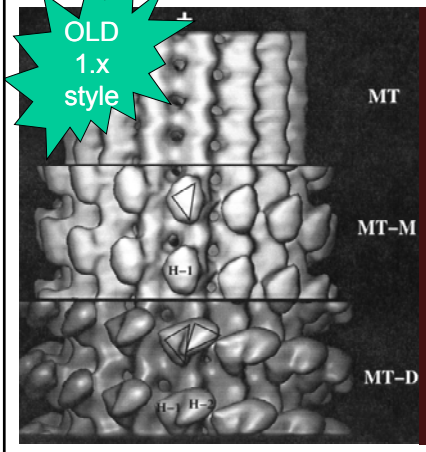
- Estimate optimum k with variability criterion.
- Index map I: $m \rightarrow n$ ($m, n = 1,\ldots,k$).
- $k! = k\,(k\text{-}1)\ldots 2$ possible combinations.
- For each index map I perform a least squares fit of the $w_{I(j)}^{(k)}$ to the $w_j^{(l)}$.
- Quality of I: residual rms deviation

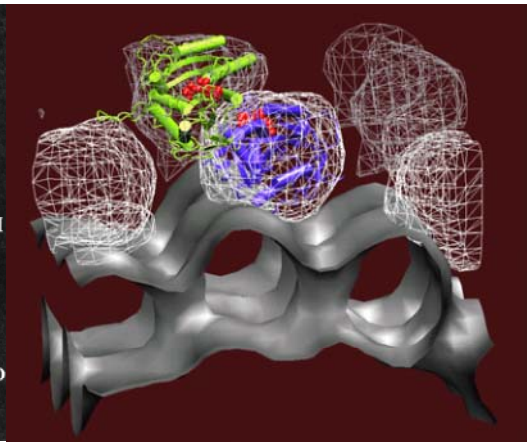
$$\Delta_I = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\left\| w_{I(j)}^{(k)} - w_j^{(l)} \right\|^2}$$

- Find optimal I by direct enumeration of the $k!$ cases (minimum of Δ_I).

Application Example: Decorated MT

OLD 1.x style

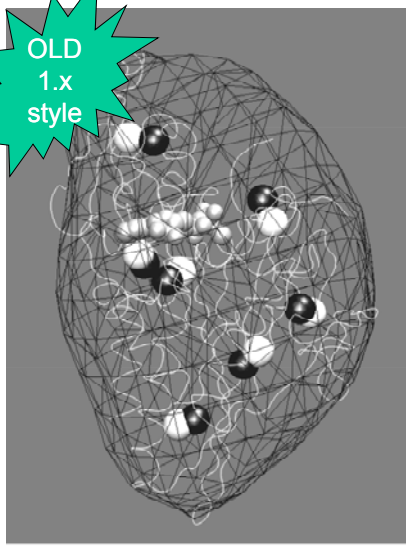




ncd monomer and dimer-decorated microtubules (Milligan *et al.*, 1997)
ncd monomer crystal structure (Fletterick *et al.*, 1996,1998)

Search for Conformations

OLD 1.x style



Two possible ranking criteria:

- Codebook vector rms deviation (Δ_I).
- Overlap between both data sets:

  Correlation coefficient:

$$C_M = \frac{\sum_{x,y,z} h_{x,y,z} \cdot l_{x,y,z}}{\left(\sum_{x,y,z} h^2_{x,y,z}\right)^{\frac{1}{2}} \left(\sum_{x,y,z} l^2_{x,y,z}\right)^{\frac{1}{2}}}$$

ncd motor (white, shown with ATP nucleotide) docked to EM map (black) using k=7 codebook vectors

Reduced Search Features

OLD 1.x style

Top 20, 7!=5040 possible pairs of codebook vectors.

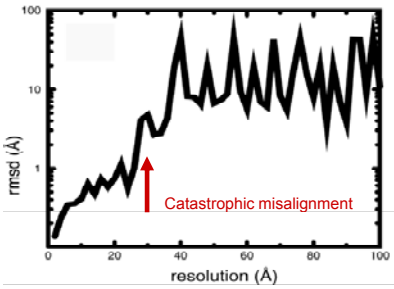
	Δ_I	C_M	I (permutation)
1.	3.115	0.913	(7,5,1,6,4,2,3)
2.	4.946	0.904	(2,3,5,7,4,6,1)
3.	5.455	0.897	(6,1,3,2,4,7,5)
4.	6.316	0.882	(5,7,4,3,1,2,6)
5.	7.612	0.867	(5,7,1,4,6,3,2)
6.	7.855	0.888	(3,2,4,1,5,6,7)
7.	7.994	0.884	(1,6,4,5,3,7,2)
8.	8.001	0.863	(6,1,4,3,5,2,7)
9.	8.192	0.888	(2,6,4,3,1,7,5)
10.	8.244	0.850	(7,5,6,2,1,3,4)
11.	8.298	0.881	(2,6,7,5,1,3,4)
12.	8.340	0.894	(6,2,4,1,3,5,7)
13.	8.481	0.867	(3,4,6,2,1,5,7)
14.	8.516	0.885	(2,3,4,5,1,7,6)
15.	8.532	0.857	(7,5,4,1,3,6,2)
16.	8.985	0.861	(6,1,5,7,4,3,2)
17.	8.988	0.838	(3,4,5,7,1,2,6)
18.	9.092	0.839	(3,2,5,4,7,1,6)
19.	9.124	0.858	(7,5,3,2,4,1,6)
20.	9.236	0.858	(1,6,5,7,4,2,3)

For a fixed k, codebook rmsd is more stringent criterion than correlation coefficient!

Performance (I)

OLD 1.x style

Dependence of accuracy on resolution (simulated EM map, automatic assignment of k from $3 \leq k \leq 9$) with Situs *qrange* tool.



Deviation from start structure (PDB: 1TOP) used to generate simulated EM map.

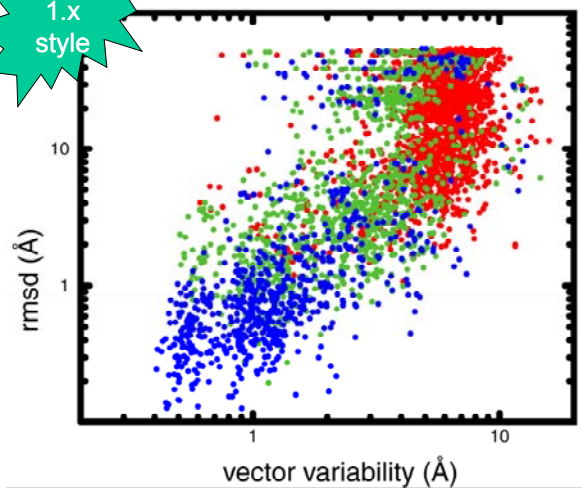
Accurate matching up to ~30Å

Performance (II)

OLD 1.x style

Is minimum vector variability a suitable choice for optimum k?

Wriggers & Birmanns, J. Struct. Biol 133, 193-202 (2001)



10 test systems, $3 \leq k \leq 9$ simulated EM densities from 2-100Å.

2-20Å (reliable fitting)
22-50Å (borderline)
52-100Å (mismatches)

Reasonable correlation with actual deviation

No "false positives" for resolution values < 20Å and variability < 1Å.

OLD 1.x style

qrange (Situs 1.x)

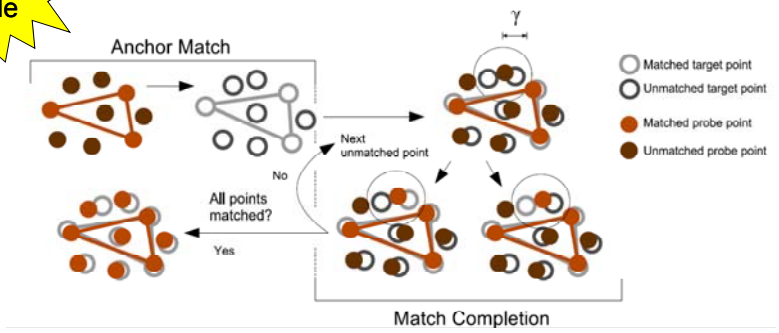
Advantages:
•Fast (seconds of compute time).
•Reduced search is robust.

Limitations:
•Original $k \rightarrow k$ algorithm for limited $\mathbf{3 \leq k \leq 9}$ works best for single molecules, not for matching subunits to larger densities.

•*matchpoint* in Situs 2.5 allows $k \rightarrow h \neq k$ matching

Anchor Point Registration: *matchpoint*

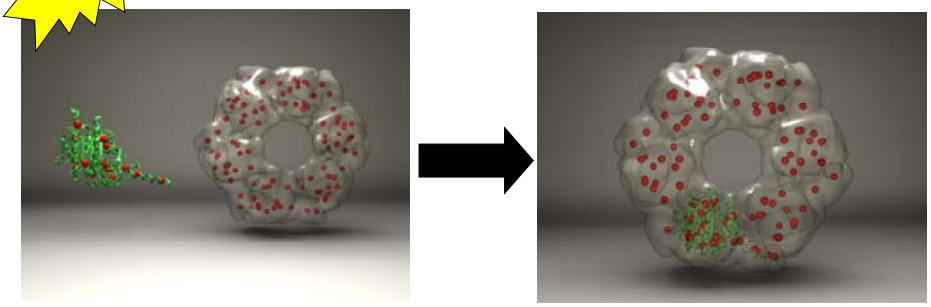
current style



Birmanns & Wriggers *J. Struct. Biol.* (2007) 157:271

Anchor Point Registration
New in Situs 2.5: *matchpoint*

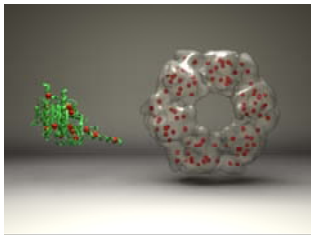
current style



- $k \rightarrow h \neq k$ matching
- number of points k (atomic), h (EM) now determined by desired level of detail, not "variability criterion". k and h should give similar point density and are dependent on volume of atomic structure and EM map

How to Determine Number of Points?

current style



Also relevant for flexible fitting (below)!

- Divide volume of EM map by volume of a "resolution element" (cube with dimension of numeric resolution value in Å).
- This gives the (maximum) number of resolved spatial features in the map.
- To avoid overfitting, we typically pick 50% of that maximum number for h.
- k is then h times the ratio of atomic to EM volume (yielding same point density, i.e. level of detail, as EM coarse graining).
- note that spatial resolution of coarse grained model scales with cubic root of number of points, so order of magnitude estimate for number of EM points h is sufficient, but k/h must closely reflect the atomic to EM volume ratio to be consistent.

Clarification of Paradigm Shift in Coarse–Graining Currently Under Way

Old style: Situs 1.x, *qrange*	Current style: Situs 2.5, *matchpoint*, *qplasty*
rigid body docking only	rigid-body docking and flexible docking (see below)
restricted to single molecule matching, i.e. $h = k$	OK to dock subunits into larger EM maps of assemblies, i.e. $h > k$
limited range $3 \leq k \leq 9$ explored automatically by exhaustive enumeration of $k!$ possibilities	no restriction on number of points (tree pruning in *matchpoint* handles larger number efficiently)
number of points typically selected from this limited range by using lowest variability criterion (justification: see slide "Perfomance II" above)	number of points now typically estimated based on number of resolved features (volume / EM resolution analysis in previous slide)

Take-Home Messages

Rigid body docking precision about one order of magnitude above the nominal EM (and SAXS) resolution

situs.biomachina.org
(UNIX command-line tools)

sculptor.biomachina.org
(GUI-based program)